Data Analysis and Visualization in R (IN2339)

Exercise Session 6 - Graphically supported hypotheses

Daniela Klaproth-Andrade, Felix Brechtmann, Julien Gagneur

Section 00 - Getting ready

1. Make sure you have already installed and loaded the following libraries:

```
library(ggplot2)
library(data.table)
library(magrittr)   # Needed for %>% operator
library(tidyr)
```

Section 01 - Color guidelines

What are best practices when using color for data visualizations? Select all that apply.

1. Avoid having too many colors for categorical data.
2. Use one bright color to attract the readers attention.
3. Use color only when it actually adds meaning to the plot.
4. Use divergent color scales for categorical data types.

Section 02 - Cofounding factors

Investigate the file `coffee_sim.csv` by first loading it as a `data.table`.

```
coffee_dt <- fread("./extdata/coffee_sim.csv")
coffee_dt
summary(coffee_dt)
```

1. Visualize the trend between coffee and coronary heart disease (CHD)-related deaths (risk), which suggests a possible causal relationship.

2. From this plot you could conclude that coffee causes CHD. Do you think this conclusion explains the original observation? Provide plots supporting other conclusions.

Section 03 - Supporting hypotheses with visualizations

1. Read the `titanic.csv` file into a `data.table`. You can read the description of the dataset on kaggle: https://www.kaggle.com/c/titanic/data.

2. Describe what you see in the data. Have a look at the first and last observations. Make a summary of the variables in the dataset.

3. What do you think are the factors that have the strongest influence on the survival rate? Make claims and justify your argument with plots. *Hint*: check variables like `pclass`, `sex` and `age`, and visualize whether they associate with survival. Additionally check their interactions.

Section 04 - General guidelines in data visualization

Below is a graph taken from one published paper. Read the figure legend.

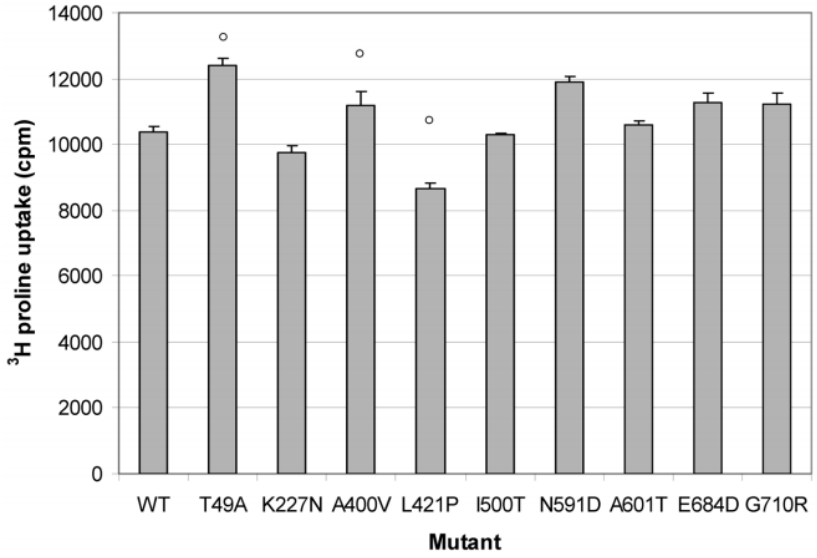

Figure 2. Maximal ^{3}H proline uptake of wildtype (WT) and all tested mutants. The maximum in uptake was measured in the presence of 3 µM cold L-proline. Data are expressed as means ± standard deviation (SD) obtained from triplicate samples. Mutants with a circle were tested in a second independent experiment.
doi:10.1371/journal.pone.0068645.g002

1. [OPTIONAL] Discuss good and bad graphical properties of the plot, make suggestions on how to improve.
2. [OPTIONAL] Implement a better visualization. As the original data is not available, we use the data simulated with the code below.

```
# simulate data
dt <- data.table(pro_uptake = c(rnorm(3, 10100, 300), rnorm(4, 12100, 300),
                                rnorm(3, 9850, 300), rnorm(4, 11100, 300),
                                rnorm(4,8300, 300), rnorm(3,10050, 300),
                                rnorm(3, 12000, 300), rnorm(3, 10020, 300),
                                rnorm(3, 10080, 300), rnorm(3, 10070, 300)  ),
                 mutants = c(rep('WT',3), rep('T49A',4), rep('K227N',3), rep('A400V',4),
                             rep('L421P',4), rep('I500T',3), rep('N591D',3),
                             rep('A601T',3), rep('E684D',3), rep('G710R',3) )          )
```

Section 5 - Case Study Feedback

Feedback for the report (Rmd) with the entire analysis

- ☐ Does the notebook run and create all figures from the presentation?
- ☐ Is the notebook cleaned-up?
- ☐ Is the report (Rmd) stand alone? (Only Rmd needed to understand the performed analysis. It should contain explanations/interpretations and code.)
- ☐ Is the data after the data preparation tidy? Is the definition of an observation clear?

Feedback on the presentation / slides

General considerations

- ☐ Has the presentation a clear structure?
- ☐ Did the presentation tell a clear and convincing story?
- ☐ Did the presenter stick to the 7 min limit?

Introduction

- ☐ Did the presentation start with a short motivation?
- ☐ Are the goals of the analyses formulated at the beginning of the presentation?

Data Preparation

- ☐ Were important data preparation steps explained during the presentation?
- ☐ Was additional data used in the presentation? If, was it made clear how that data was obtained?

Data Exploration/Analysis

- ☐ Were the stated claims communicated well?
- ☐ Were all stated claims supported by appropriate figures?
  - Was the appropriate plot type (line plot, scatter plot, box plot, violin plot) selected for the data shown?
  - Did the plot support the claim? Is there a better alternative to visualize the claim?
  - Was an associative plot used to show the relationship stated in the claim.
- ☐ Did the figures follow the plotting guidelines (no double-encoding, good paper-ink-ratio)?
- ☐ Were alternative interpretations of the associations discussed (directionality, effect of a third variable, robustness)?
- ☐ Where there any circularities in the presented claims?

Conclusion

- ☐ Did the presentation end with a conclusion slide recapping the main findings?
- ☐ Did the claims answer the problem/goals formulated in the motivation?
- ☐ Did the presentation help to learn something that we/you did not know before? (non-evident claims/association/pattern/relationship in the data presented)